Instructions: The is a closed-notes, closed-book exam. No calculator or electronic device of any kind may be used. Use nothing but a pen or pencil. Please write your name and answers on blank paper. Please do NOT write your answers on the pages with the questions. For questions that require extensive numerical calculations that you should not be expected to do without a calculator, simply set up the calculation and leave it at that. For example, $(3.45 - 1.67)/\sqrt{2.34}$ would be an acceptable answer. On the other hand, some quantities that are very difficult to compute one way may be relatively easy to compute another way. Part of this exam tests your ability to figure out the easiest way to compute things, based on the information provided and the relationships between various quantities. In general, I expect you to be able to invert diagonal matrices of any dimensions and non-diagonal 2×2 matrices. If you find yourself trying to deal with something more complicated, there is probably a better way to solve the problem.

1. Consider an experiment conducted at two research labs. Within each lab, four mice were assigned two treatments using a completely randomized design with two mice per treatment. Let y_{ijk} be the response variable measurement for the kth mouse that received treatment j in research lab i ($i = 1, 2$; $j = 1, 2$; $k = 1, 2$). Suppose

$$y_{ijk} = \mu + \lambda_i + \tau_j + \epsilon_{ijk}, \tag{1}$$

where μ, λ_1, λ_2, τ_1, and τ_2 are unknown parameters and the ϵ_{ijk} terms are independent normal random variables with mean 0 and some unknown variance σ^2. Let

$$\boldsymbol{y} = (y_{111}, y_{112}, y_{121}, y_{122}, y_{211}, y_{212}, y_{221}, y_{222})',$$

$$\boldsymbol{\epsilon} = (\epsilon_{111}, \epsilon_{112}, \epsilon_{121}, \epsilon_{122}, \epsilon_{211}, \epsilon_{212}, \epsilon_{221}, \epsilon_{222})',$$

and

$$\boldsymbol{\beta} = (\mu, \lambda_1, \lambda_2, \tau_1, \tau_2)'.$$

   (a) Give the entries in a matrix $\boldsymbol{X}$ so that the model defined in equation (1) above can be written as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

   (b) Give the entries in a matrix $\boldsymbol{X}^*$ that has all of the following properties:
      - $\boldsymbol{X}^*$ has the same column space as $\boldsymbol{X}$.
      - $\boldsymbol{X}^*$ has full-column rank.
      - The columns of $\boldsymbol{X}^*$ are orthogonal; i.e., if $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$ are any two columns of $\boldsymbol{X}^*$, then $\boldsymbol{x}_u'\boldsymbol{x}_v = 0$.

2. Consider an experiment with two factors: A (with levels A_1 and A_2) and B (with levels B_1 and B_2). Suppose an unbalanced, completely randomized design was used to assign experimental units to treatments. For $i = 1, 2$ and $j = 1, 2$, let n_{ij} be the number of experimental units treated with level A_i of factor A and level B_j of factor B. Let y_{ijk} denote the response of the kth experimental unit treated with level A_i of factor A and level B_j of factor B for $i = 1, 2$, $j = 1, 2$, and $k = 1, \ldots, n_{ij}$. Assume that

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \tag{2}$$

where each μ_{ij} term is an unknown mean parameter and all ϵ_{ijk} terms are independent and identically distributed as $N(0, \sigma^2)$ for some unknown variance parameter σ^2. The following table contains the observed response data. (Whole numbers are presented to make calculation easier. Individual observations are separated by commas within each cell of the table. Samples sizes are $n_{11} = 2$, $n_{12} = 3$, $n_{21} = 1$, and $n_{22} = 2$.)

Observed Response Values (y_{ijk} values)

	B_1	B_2
A_1	3, 5	7, 8, 9
A_2	13	2, 4

(a) Provide the numerical value of the estimator you would use to estimate σ^2.

(b) Find an LSMEAN for each level of factor A.

(c) Find the Type III sum of squares for factor A in the Type III ANOVA table for the full model that includes an intercept, a factor A main effect, a factor B main effect, and $A \times B$ interaction.

(d) Suppose a researcher attempts to fit a reduced model with no factor B main effect using the R code below.

```
> y=c(3,5,7,8,9,13,2,4)
> A=factor(c(1,1,1,1,1,2,2,2))
> B=factor(c(1,1,2,2,2,1,2,2))
> o=lm(y~A+A:B)
> model.matrix(o)
  (Intercept) A2 A1:B2 A2:B2
1           1  0     0     0
2           1  0     0     0
3           1  0     1     0
4           1  0     1     0
5           1  0     1     0
6           1  1     0     0
7           1  1     0     1
8           1  1     0     1
```

If the R command `summary(o)` were issued next, the researcher would see parameter estimates labeled `Intercept`, `A2`, `A1:B2`, and `A2:B2`, along with standard errors for those estimates. Provide those estimates and their standard errors.

(e) Suppose the researcher continues analysis by issuing the following R commands:

```
> #Remove the second column from
> #the previous model matrix.
> X=model.matrix(o)[,-2]
> #Fit a model with X as the model matrix.
> o=lm(y~0+X)
> model.matrix(o)
  X(Intercept) XA1:B2 XA2:B2
1            1      0      0
2            1      0      0
3            1      1      0
4            1      1      0
5            1      1      0
6            1      0      0
7            1      0      1
8            1      0      1
```

This code fits a model that is reduced relative to the cell means model in equation (2). Explain what this reduced model implies about μ_{11}, μ_{12}, μ_{21}, and μ_{22}.

3. Consider a balanced and completely randomized experimental design with four treatments and 10 experimental units per treatment (40 experimental units in total). Suppose y_{ij} represents the response variable measurement for the jth experimental unit treated with treatment i ($i = 1, 2, 3, 4; j = 1, \ldots, 10$). Assume that

$$y_{ij} = \mu_i + \epsilon_{ij}, \tag{3}$$

where each μ_i term is a mean parameter and all ϵ_{ij} terms are independent and identically distributed as $N(0, \sigma^2)$ for some unknown variance parameter σ^2.

(a) Find a simplified, non-matrix expression for the F-statistic you would use to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

(b) State the degrees of freedom for the F-statistic provided in part (a).

(c) Suppose $\mu_1 = 9$, $\mu_2 = 7$, $\mu_3 = 5$, $\mu_4 = 5$, and $\sigma^2 = 2.5$. Determine the numerical value of the non-centrality parameter for the F-statistic in part (a). (If done carefully, you can compute this quantity exactly without a computer or calculator.)

4. Researchers were studying the effect of fertilizer on crop yield. Fertilizer amounts $x_1 = 0$, $x_2 = 20$, $x_3 = 40$, $x_4 = 60$, $x_5 = 80$, and $x_6 = 100$ units were randomly assigned to a total of 12 plots of land, with 2 plots per fertilizer amount. Let y_{ij} be the yield for the jth plot that received x_i units of fertilizer ($i = 1, \ldots, 6; j = 1, 2$). Assume that

$$y_{ij} = \mu_i + \epsilon_{ij}, \tag{4}$$

where each μ_i is an unknown parameter and all ϵ_{ij} terms are independent and identically distributed as $N(0, \sigma^2)$ for some unknown variance parameter σ^2. In the R code below y is the response vector

$$\boldsymbol{y} = (y_{11}, y_{12}, y_{21}, y_{22}, y_{31}, y_{32}, y_{41}, y_{42}, y_{51}, y_{52}, y_{61}, y_{62})'.$$

```
> x=c(0,0,20,20,40,40,60,60,80,80,100,100)
> o=lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5))
> anova(o)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x          1 7875.0  7875.0 289.8773 2.626e-06 ***
I(x^2)     1 1314.9  1314.9  48.4005 0.0004379 ***
I(x^3)     1   62.5    62.5   2.3006 0.1801125
I(x^4)     1   64.3    64.3   2.3663 0.1748964
I(x^5)     1   56.0    56.0   2.0613 0.2010907
Residuals  6  163.0    27.2
```

(a) A special case of model (4) assumes there exist β_0, β_1, and $\beta_2 \in \mathbb{R}$ so that $\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$ for all $i = 1, \ldots, 6$. Provide a test statistic that can be used to test whether this special case of model (4) fits the data adequately.

(b) State the null distribution of the test statistic in part (a).

(c) A researcher believes that the following special case of model (4) holds. Mean yield will increase linearly as the fertilizer amount increases from 0 to 50 units, and mean yield will remain constant and equal to the mean yield for 50 units of fertilizer for any fertilizer amount from 50 to 100 units. Provide a matrix $\boldsymbol{X}$ that can serve as the model matrix for the researcher's proposed model.

(d) Suppose the sum of squared errors is 305.4 for the model described in part (c). Does this model fit the data adequately relative to model (4)? Provide a test statistic that can be used to answer this question.

(e) State the degrees of freedom associated with the statistic provided in part (d).